# Social Network Research in the Age of Computation

Mohammad Mahdian

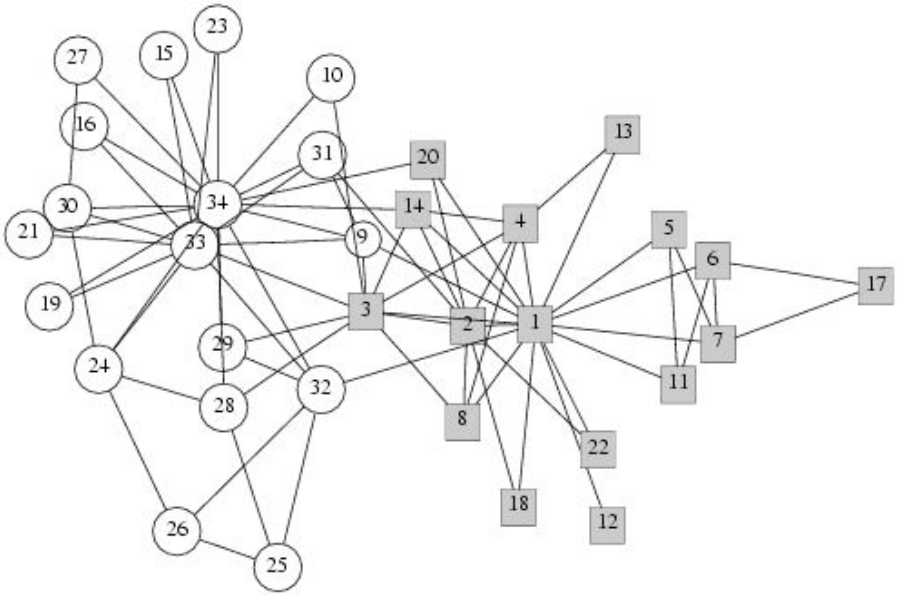
Yahoo! Research

# What is a social network?

- Social network: a graph that represents pair-wise interactions among a group of individuals/ independent entities.
- provides an abstraction of the structure and dynamics of diverse kinds of interaction.
- SNs are everywhere, and have been around forever
  - Friendship
  - Sexual relationship
  - Scientific collaborations
  - IM contacts
  - …

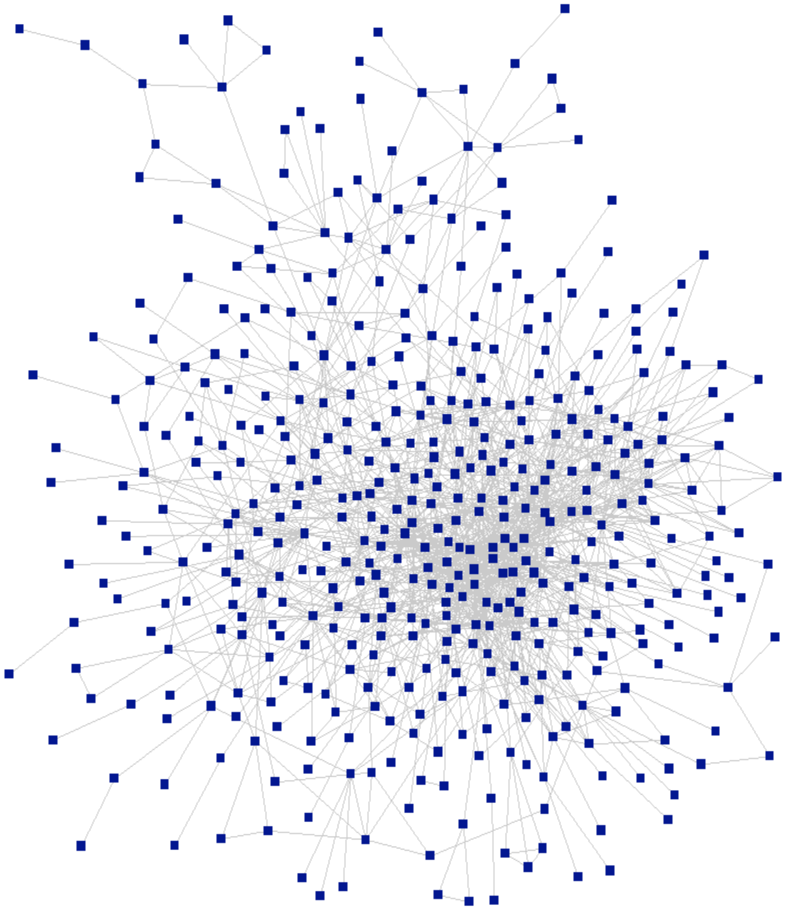
# Example: friendships in a karate club

- Wayne Zachary (1977) recorded friendships among 34 members of a karate club at a university over 2 years.



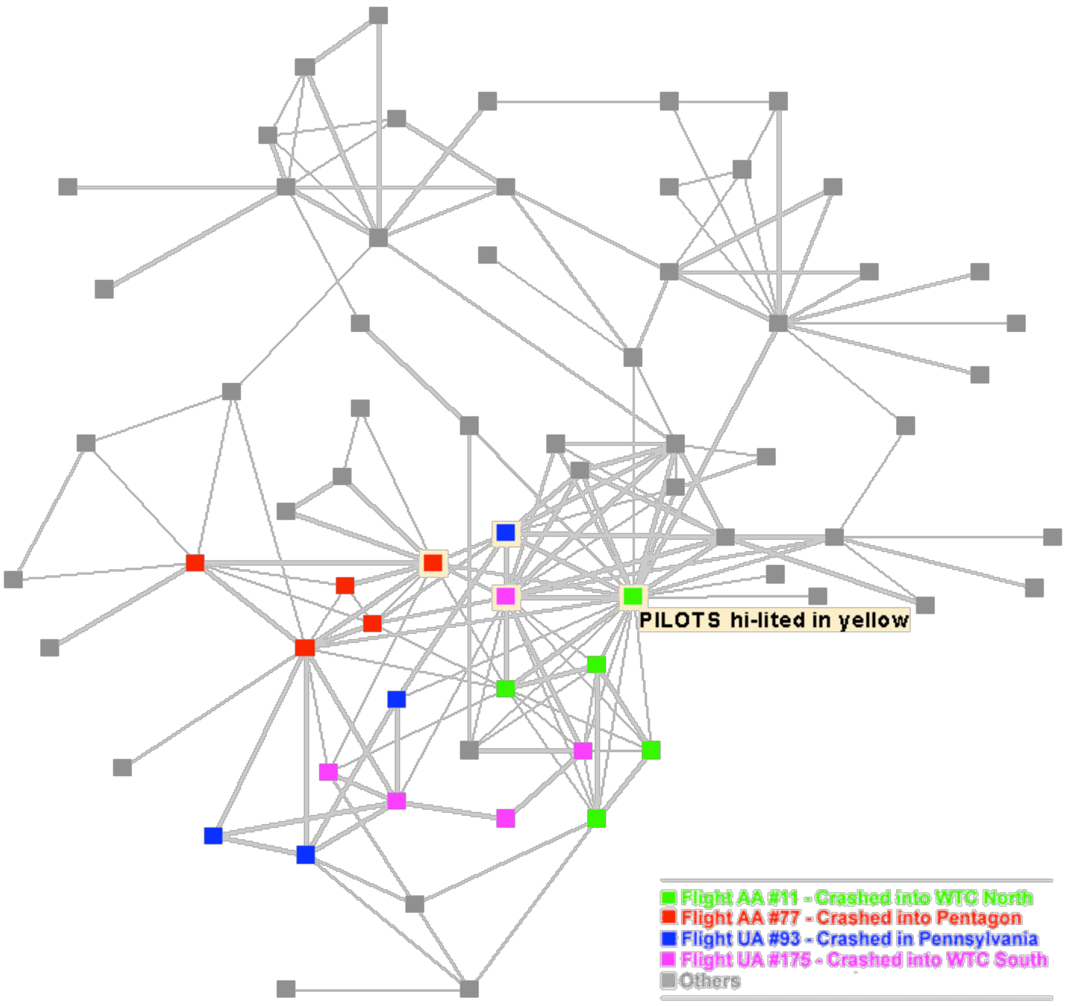
# Example: Scientific Collaboration

- 400,000 nodes, authors in *Math Reviews* DB
- edge between two authors if they have a joint paper
- ~ 676,000 edges
- Many low-degrees (100K of deg 1), few high-degs (509, 268, 244, …)



Picture from orgnet.com

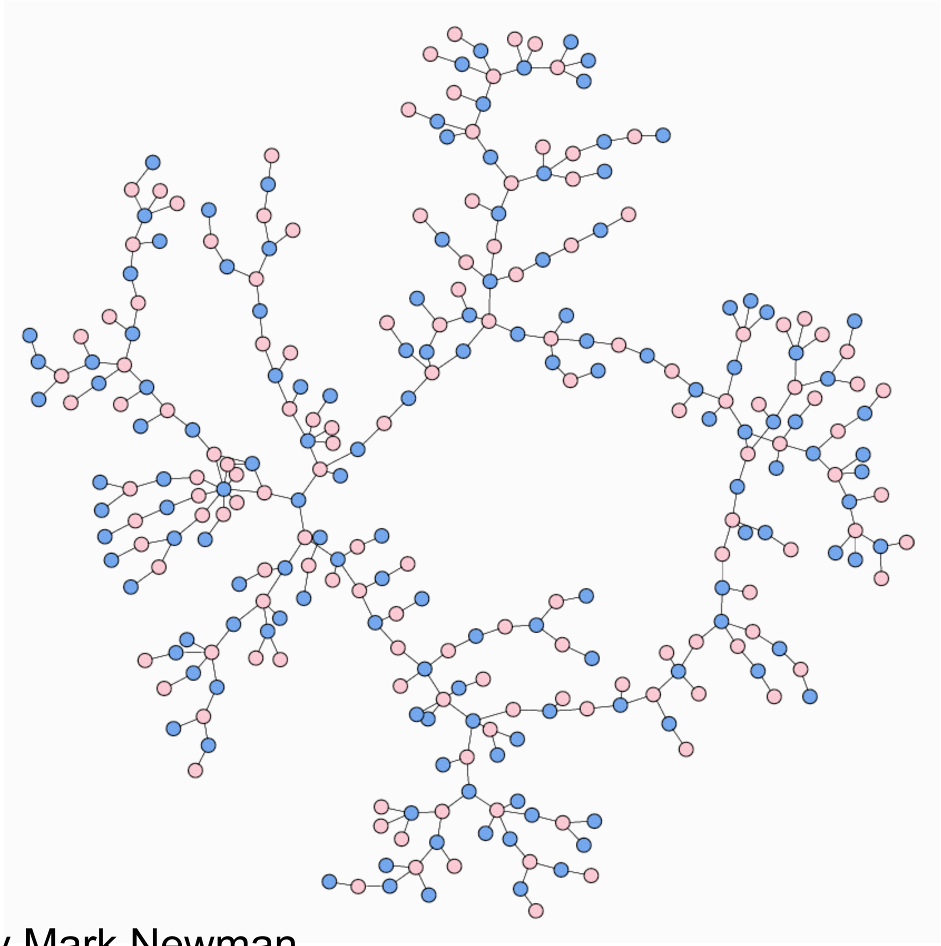
# Example: 9/11 Terrorist Network


PILOTS hi-lited in yellow
🟩 Flight AA #11 - Crashed into WTC North
🟥 Flight AA #77 - Crashed into Pentagon
🟦 Flight UA #93 - Crashed in Pennsylvania
🟪 Flight UA #175 - Crashed into WTC South
⬜ Others

Picture from orgnet.com

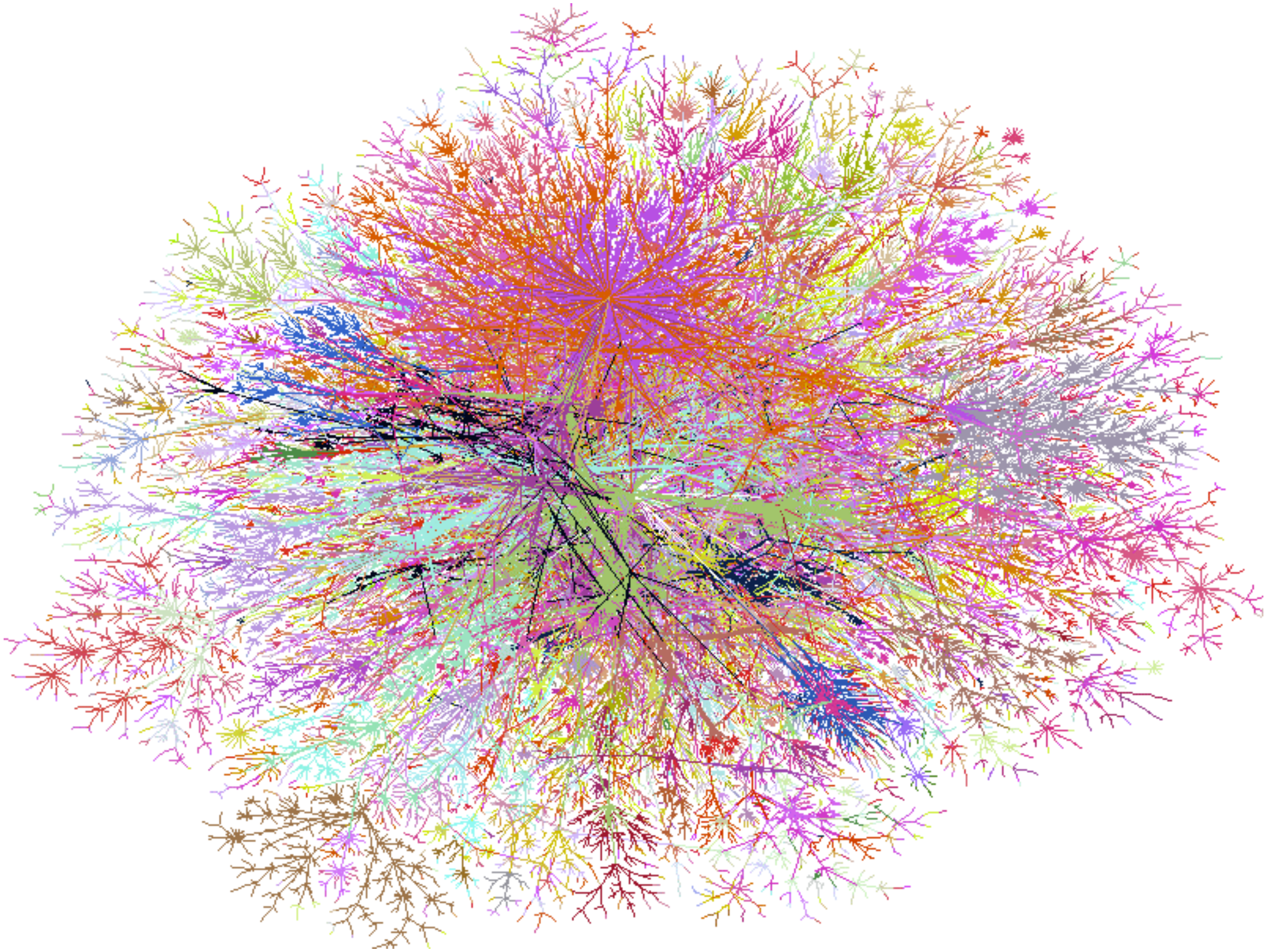
# Example: high-school dating

- Data collected through in-school questionnaires and in-house interviews at a high-school in a midwestern town



Bearman, Moody, and Stovel; picture by Mark Newman

# Example: The Internet



# So, what’s new?

- More and more, “interactions” are moving to the digital world
  - Either people interacting digitally (e.g., on the web), or leaving digital traces
- So, we are collecting data on such interactions on a massive scale

) Lots of raw material for research

# What’s new, cont’d.

- More and more of today’s social systems are engineered (not “organically grown”)
  - ❑ Web 2.0 revolution
  - ❑ Large-scale distributed collaboration systems (e.g., Wikipedia)
  - ❑ Telecommunication costs

) more demand for social network research

# Aspects of modern SN research

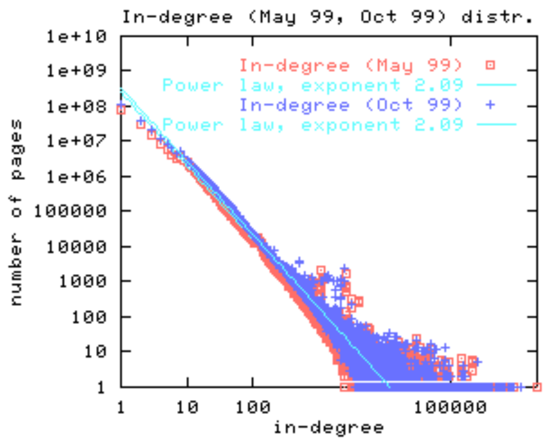
- Incentives (economics, social psychology)
- Massive data sets
  - E.g.: twitter generates about 1.8 TB/month
  - Need efficient algorithms and tools such as grid computing
- Noisy data
- Often hard to perform experiments (due to cost/privacy/commercial reasons), but can observe online users in their “natural habitat”.

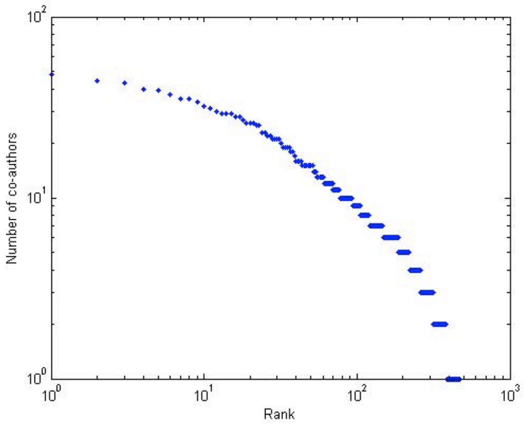
# Social Network Research

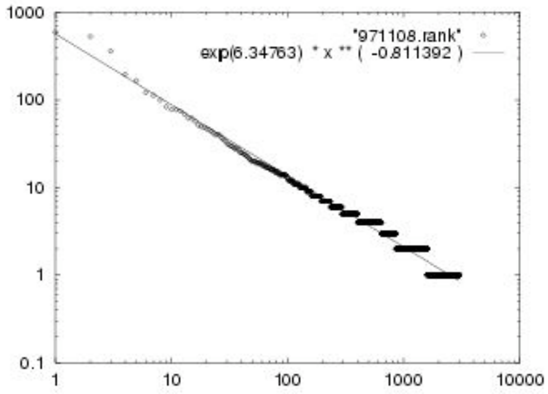
- An interdisciplinary field of research between
  - Computer science (algorithms, AI, data mining, HCI, …)
  - Sociology
  - Economics
  - Social psychology, physics, anthropology, epidemiology, …
- The goal of this field is to
  - Observe micro-level preferences and macro-level phenomena that are common in SNs
  - Propose and analyze models that explain how "micro-motives" lead to "macro-behaviors"
  - Give efficient computational methods to mine social network data

# Example: power law degree distributions

- Many SNs obey a heavy-tail/power-law degree distribution, i.e., # nodes of deg k is proportional to $k^{-c}$.







- Is there a simple model that explains this?

# Power laws and preferential attachment

- Barabasi and Albert (1999):
  - Nodes are added one by one.
  - Each new node chooses k old nodes to connect to.
  - The probability of choosing a node is proportional to its current degree.
- This process yields a power law degree distribution (Bollobas et al., 2001).
- Other similar models demonstrate that generally “The rich gets richer” phenomenon often results in heavy-tailed distributions.

# Why study social networks?

- Understanding the nature of behaviors of human individuals and societies
- Predicting possible social outcomes or influencing the outcome:
  - epidemics
    - Changes in transportation costs has structurally changed the social network of physical interactions. What does this mean for disease epidemics?
    - The effect of sexual behavior on STD prevalence (Morris et al., Am. J. of Public Health, 2009)

# Applications of SN research, cont’d

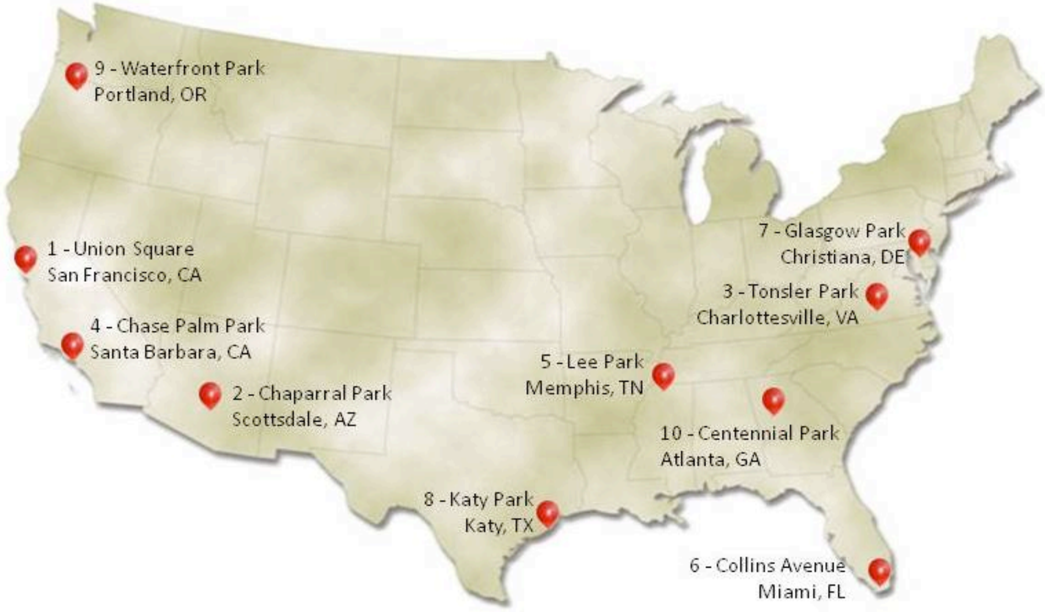
- Language evolution
  - Languages evolve as a result of human interaction
  - Can we automatically track language evolution and use this for NLP applications?
- Polarization/Balcanization of (online) societies
  - What’s the role of communication platform?
- Technology diffusion
  - Say, a new communication technology is introduced.
  - Users won’t use it unless their friends use them.
  - Marketing question: What strategies are effective in promoting the new technology?

# Applications of SN research, cont'd

- Using the power of social networks
  - Essentially a very large, capable, sensor network
  - However, nodes act in their own self interest.
  - Can we use this network?
  - DARPA network challenge:
    - 10 red balloons placed in different locations in the US
    - First team to find them all wins $40,000.

# Applications of SN research, cont'd

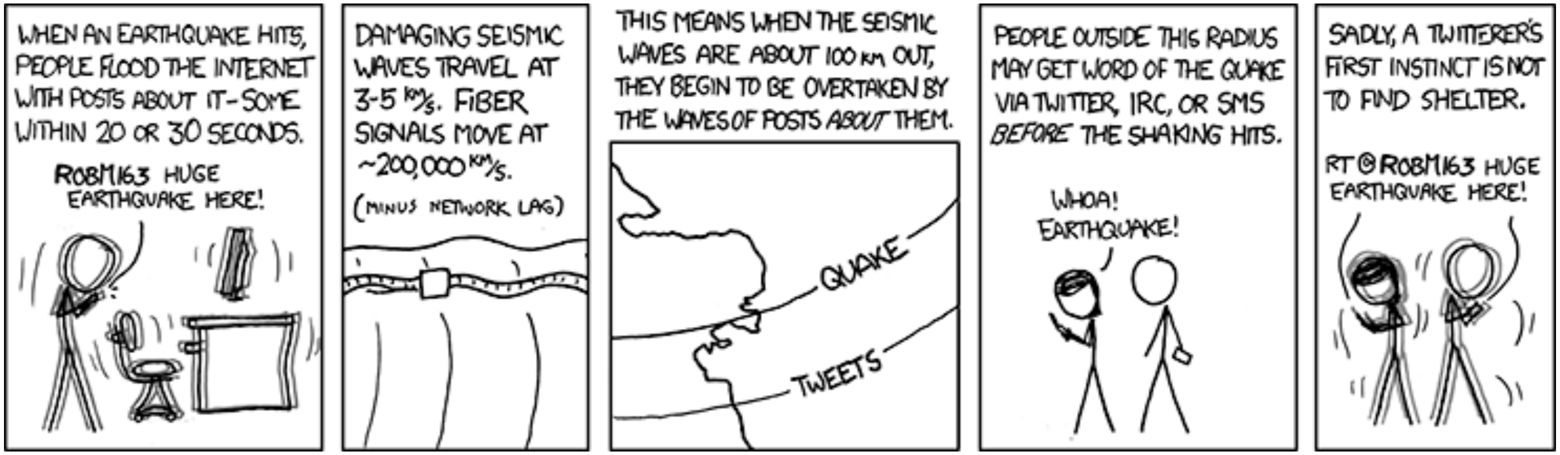
- MIT Media Lab team won



- Started a website 48 hrs before the contest
- Recruited ~5000 participants
- Found all 10 balloons in 8hrs 52mins

# Applications of SN research, cont’d

- To use the power of a social network as a “sensor” network, we must
  - Recruit agents by giving incentive
  - Figure out how to deal with incorrect data

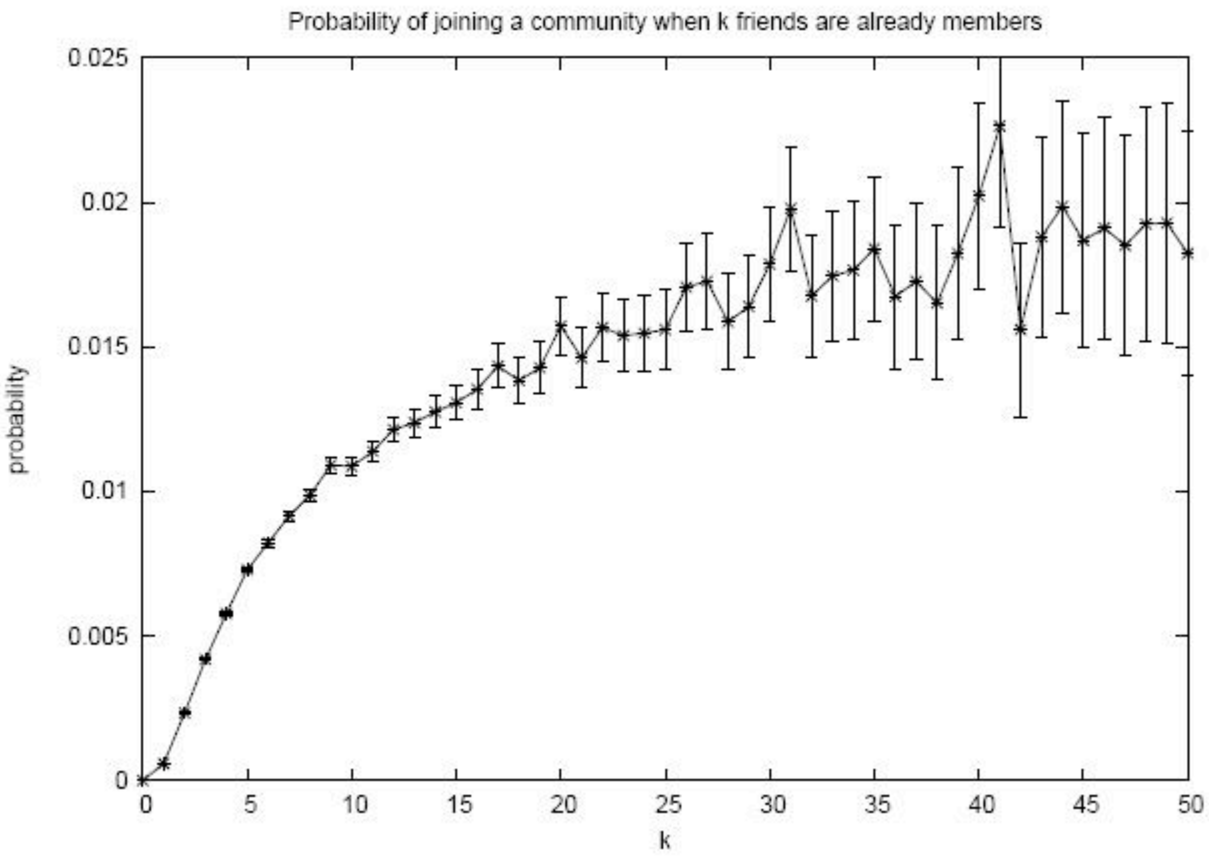

WHEN AN EARTHQUAKE HITS, PEOPLE FLOOD THE INTERNET WITH POSTS ABOUT IT—SOME WITHIN 20 OR 30 SECONDS.
ROBM163 HUGE EARTHQUAKE HERE!
DAMAGING SEISMIC WAVES TRAVEL AT 3-5 KM/S. FIBER SIGNALS MOVE AT ~200,000 KM/S.
(MINUS NETWORK LAG)
THIS MEANS WHEN THE SEISMIC WAVES ARE ABOUT 100 KM OUT, THEY BEGIN TO BE OVERTAKEN BY THE WAVES OF POSTS ABOUT THEM.
QUAKE
TWEETS
PEOPLE OUTSIDE THIS RADIUS MAY GET WORD OF THE QUAKE VIA TWITTER, IRC, OR SMS BEFORE THE SHAKING HITS.
WHOA! EARTHQUAKE!
SADLY, A TWITTERER'S FIRST INSTINCT IS NOT TO FIND SHELTER.
RT @ROBM163 HUGE EARTHQUAKE HERE!

# Social correlation

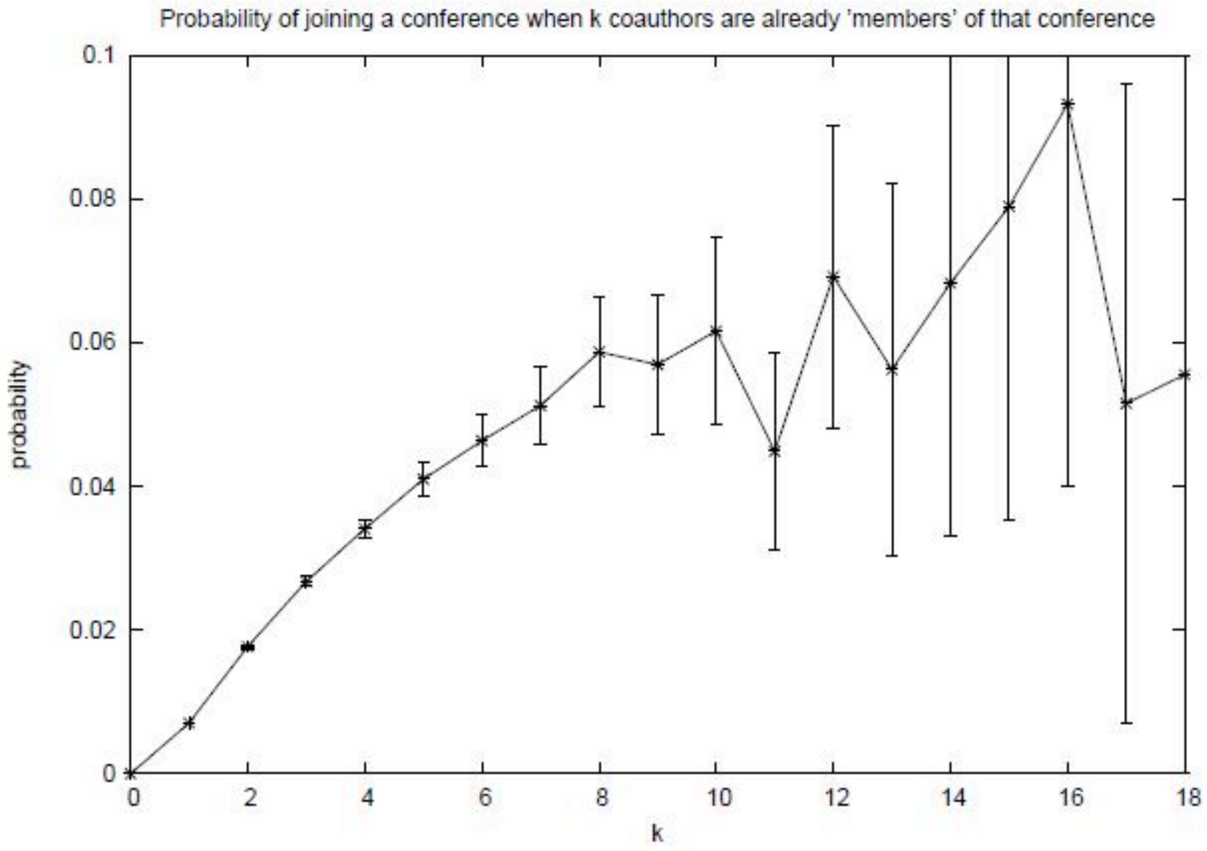
- Role of social ties in shaping the behavior of users

- Examples:
  - Joining LiveJournal communities [Backstrom et al.]
  - Publishing in conferences [Backstrom et al.]
  - Tagging vocabulary on flickr [Marlow et al.]
  - Adoption of paid VOIP service in IM
  - …

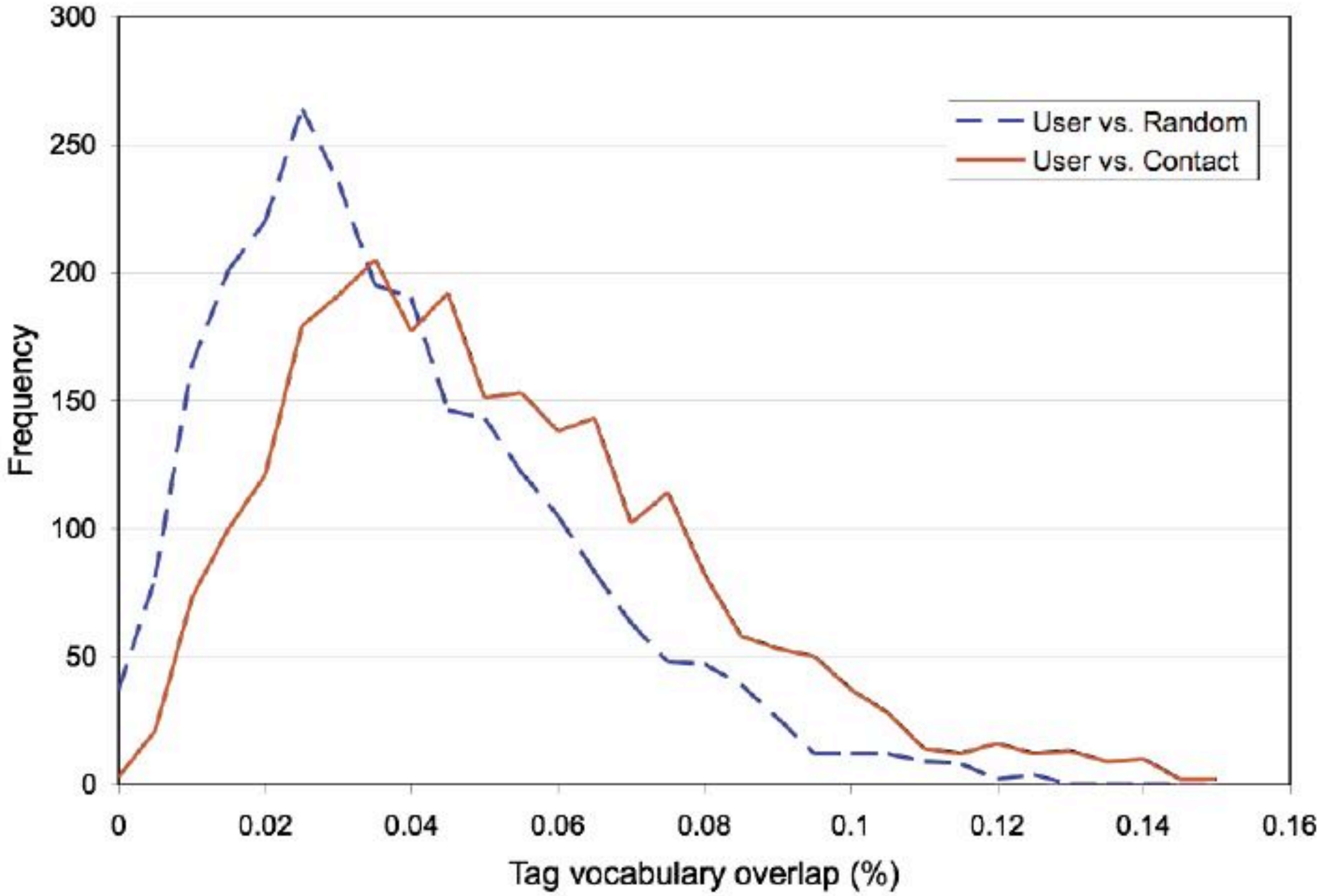
# Joining communities [Backstrom et al]



# Publishing in conferences

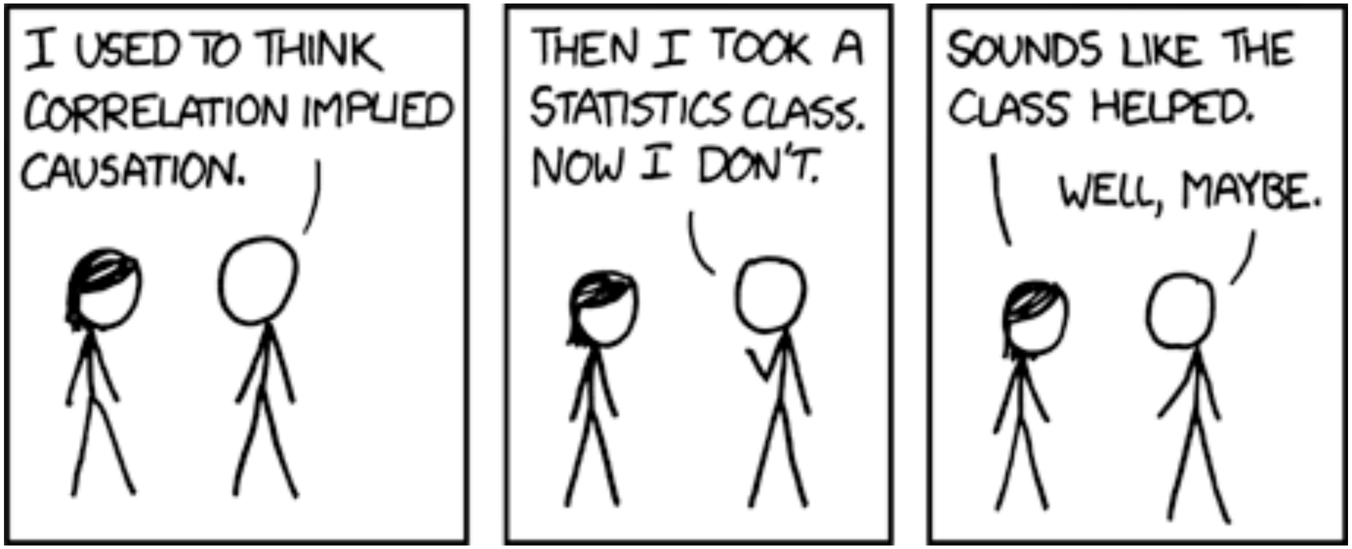


# Flickr tag vocabulary [Marlow et al.]



# Correlation vs influence

- Common mistake:  attribute the observed correlation to social influence/learning


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

# Sources of correlation

- Social influence: One person performing an action can cause her contacts to do the same.
  - by providing information
  - by increasing the value of the action to them
- Homophily: Similar individuals are more likely to become friends.
  - Example: two mathematicians are more likely to become friends.
- Confounding factors: External influence from elements in the environment.
  - Example: friends are more likely to live in the same area, thus attend and take pictures of similar events, and tag them with similar tags.

# Social influence

- Focus on a particular “action” A.
  - E.g.: buying a product, joining a community, publishing in a confernence, using a particular tag, using the VOIP service, …
- An agent who performs A is called “active”.
- x has influence over y if x performing A causes/increases the likelihood that y performs A.
- Distinguishing factor: causality relationship

# Identifying social influence

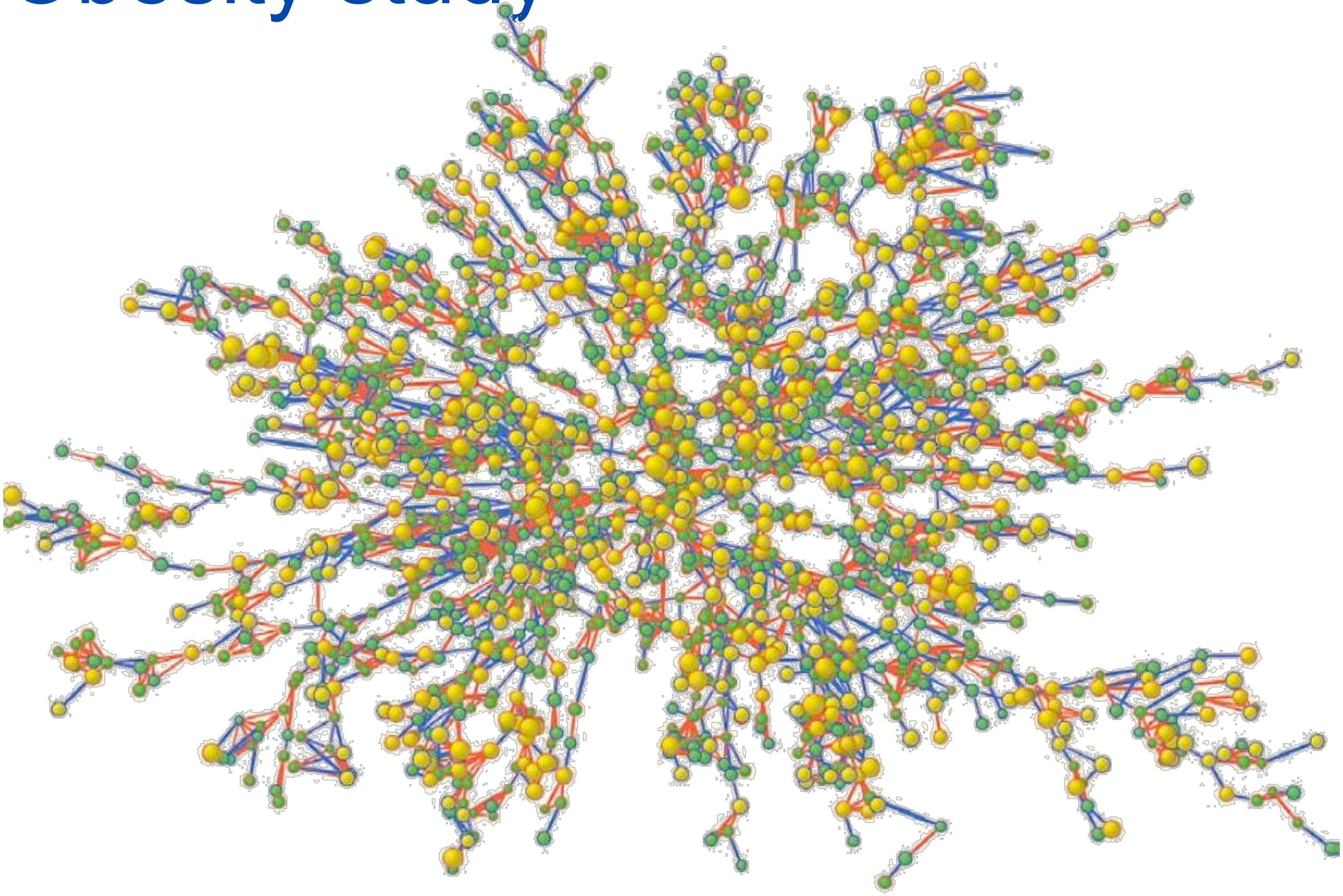
- Why is it important?
- Analysis: predicting the dynamics of the system. Whether a new norm of behavior, technology, or idea can diffuse like an epidemic.
- Design: for designing a system to induce a particular behavior, e.g.:
  - vaccination strategies (random, targeting a demographic group, random acquaintances, etc.)
  - viral marketing campaigns

# Example: obesity study

Christakis and Fowler, “The Spread of Obesity in a Large Social Network over 32 Years”, New England Journal of Medicine, 2007.

- Data set of 12,067 people from 1971 to 2003 as part of Framingham Heart Study

# Obesity study

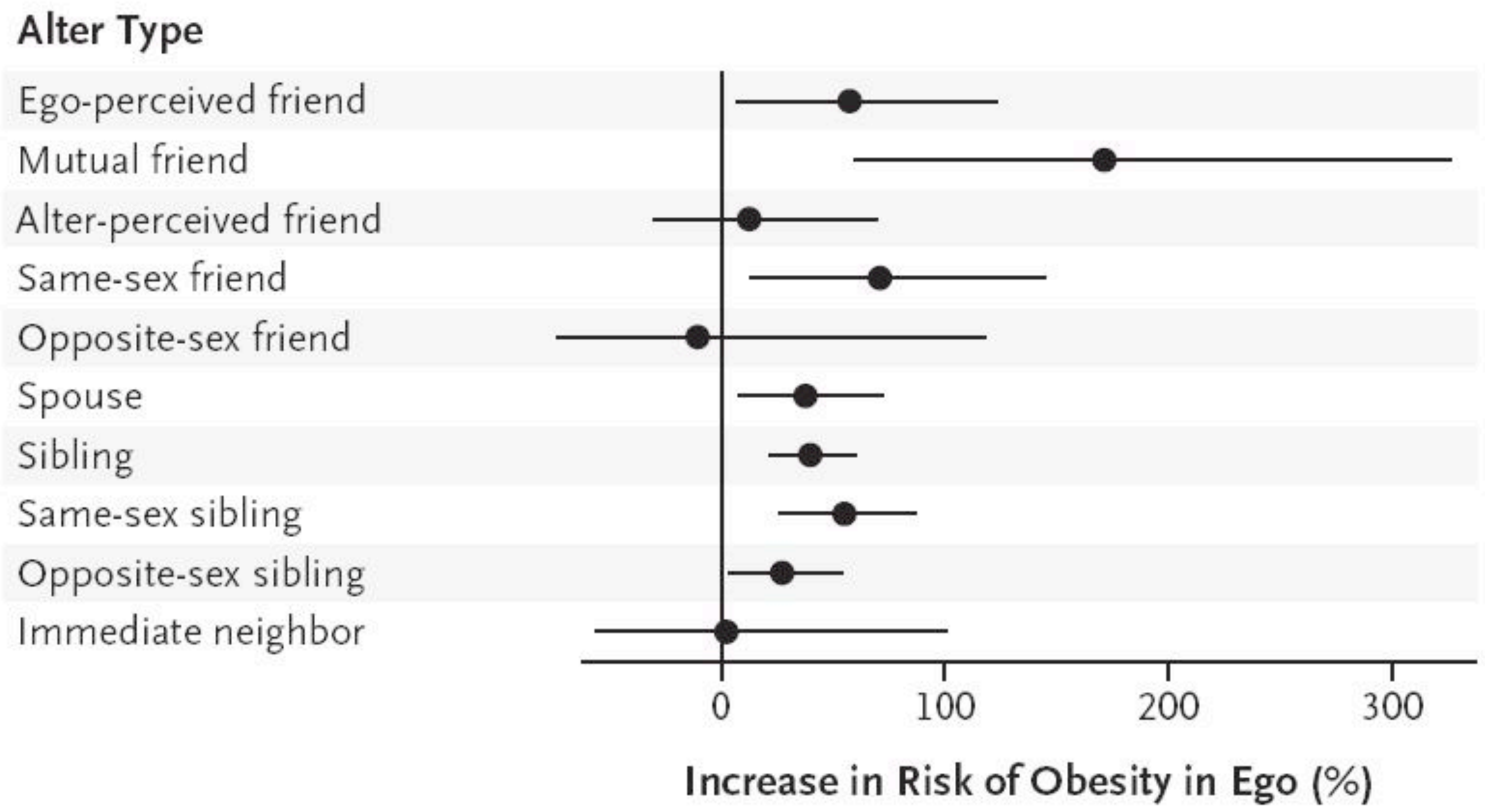


# Example: obesity study

Christakis and Fowler, “The Spread of Obesity in a Large Social Network over 32 Years”, New England Journal of Medicine, 2007.

- Data set of 12,067 people from 1971 to 2003 as part of Framingham Heart Study
- Results
  - Having an obese friend increases chance of obesity by 57%.
  - obese sibling ! 40%, obese spouse ! 37%
- Methodology
  - Logistic regression, taking many attributes into account (e.g., age, sex, education level, smoking cessation)
  - Taking advantage of data that is available over time
  - “edge reversal test”

# Obesity study



# Models of social influence

- Many models proposed in different settings
  - Game-theoretic models
    - Each agent modeled as a player in a “game”.
    - The utility that an agent derives depends on what his/her friends do.
    - Agents decide whether to become active to maximize their utility.
    - Example: adoption of a comm tech, e.g., cell-phone, IM
    - [Morris’00], [Immorlica et al.’07]
  - Probabilistic models
  - Probabilistic models
    - Independent cascade model [Kempe et al.]
      - Every neighbor u of v who becomes active gets an independent chance to influence v with probability $p_{uv}$.
    - Linear threshold model [Kempe et al.]
      - Each node has a random threshold, becomes active if sum of weights of active friends exceeds threshold.
    - Ising-type models from physics

# Models of social influence

- Probabilistic models are more predictive
  - allows optimization (find the best “seed set”)
  - allows fitting the data to estimate parameters of the system
- Our model also includes the element of time
  - Graph G; Time period [0,T]
  - At any time period a number of agents can become active
  - Let W be the set of active nodes at the end.

# Model

- Influence model: each agent becomes active in each time step independently with probability p(a), where a is the # of active friends.
- Natural choice for p(a): logistic regression function:

$$\ln\left(\frac{p(a)}{1-p(a)}\right) = \alpha \ln(a+1) + \beta$$

with ln(a+1) as the explanatory variable. I.e.,

$$p(a) = \frac{e^{\alpha \ln(a+1)+\beta}}{1+e^{\alpha \ln(a+1)+\beta}}$$

- Coefficient ® measures social correlation.

# Measuring social correlation

- We compute the maximum likelihood estimate for parameters ® and ¯.
- Let $Y_a$ = # pairs (user u, time t) where u is not active and has a active friends at the beginning of time step t, and becomes active in this step.
- Let $N_a$ = …… does not become active in this step.
- Find ®, ¯ to maximize

$$\prod_a p(a)^{Y_a}(1 - p(a))^{N_a}$$

- For convenience, we cap a at a value R.

# The max likelihood problem

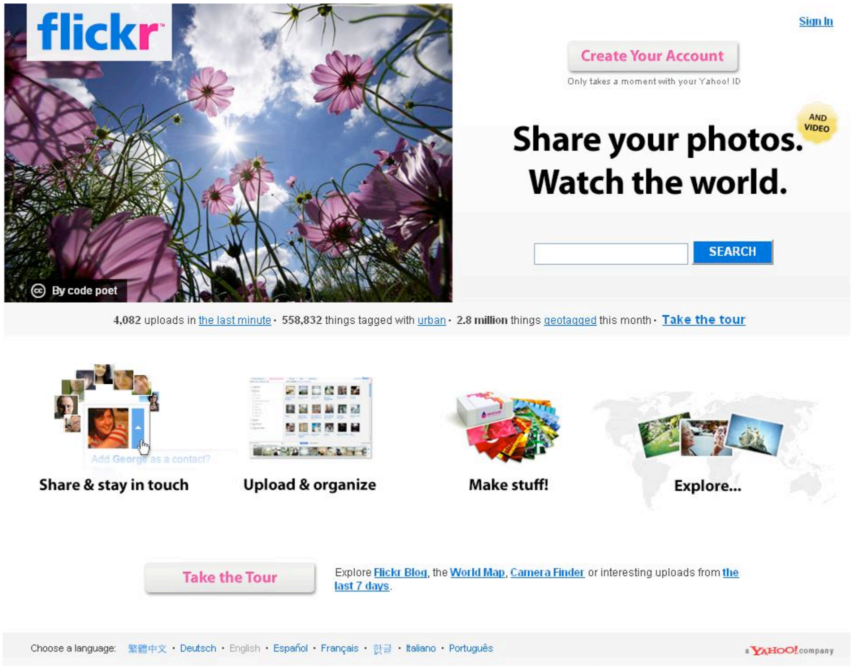
- Lemma. There is a unique solution (®, ¯) that maximizes the likelihood function.

- Proof idea. Assume (®, ¯) and (®’, ¯’) both maximize this function. We give a path between these two points such that the likelihood function is concave along this path.

- Same proof can be used to show that estimated (®, ¯) is a continuous function of $Y_a$’s and $N_a$’s.

# Flickr data set

- Photo sharing website
- 16 month period
- Growing # of users, final number ~800K
- ~340K users who have used the tagging feature
- Social network:
  - Users can specify “contacts”.
  - 2.8M directed edges, 28.5% of edges not mutual.
  - Size of giant component ~160K



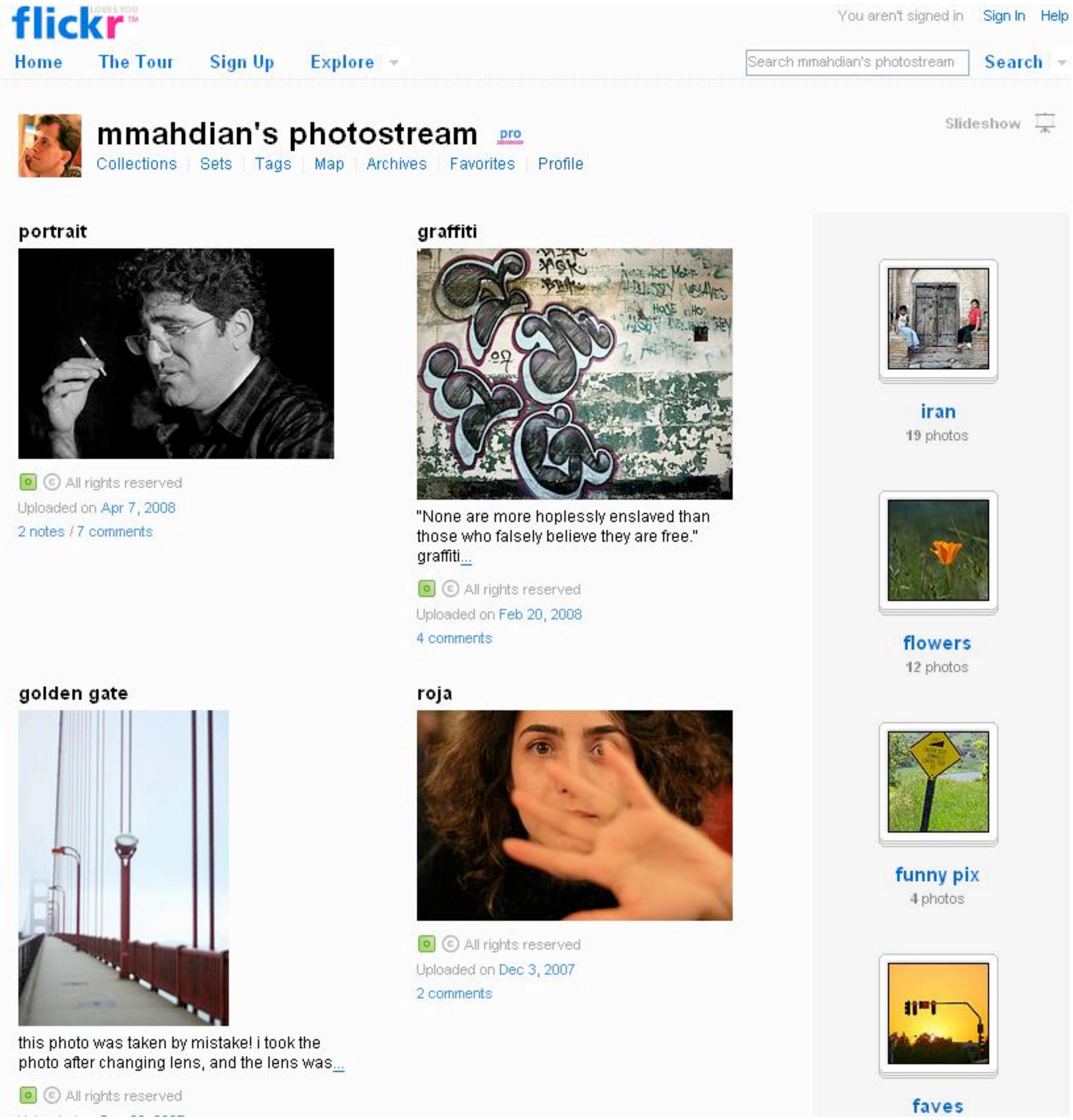
Only takes a moment with your Yahoo! ID

Share your photos. Watch the world.

4,082 uploads in the last minute · 558,832 things tagged with urban · 2.8 million things geotagged this month · Take the tour

Explore Flickr Blog, the World Map, Camera Finder or interesting uploads from the last 7 days.

Choose a language: 繁體中文 · Deutsch · English · Español · Français · 한글 · Italiano · Português



You aren't signed in Sign In Help

mmahdian's photostream pro

**portrait**

All rights reserved

Uploaded on Apr 7, 2008

2 notes / 7 comments

**graffiti**

"None are more hoplessly enslaved than those who falsely believe they are free." graffiti...

All rights reserved

Uploaded on Feb 20, 2008

4 comments

**golden gate**

this photo was taken by mistake! i took the photo after changing lens, and the lens was...

All rights reserved

**roja**

All rights reserved

Uploaded on Dec 3, 2007

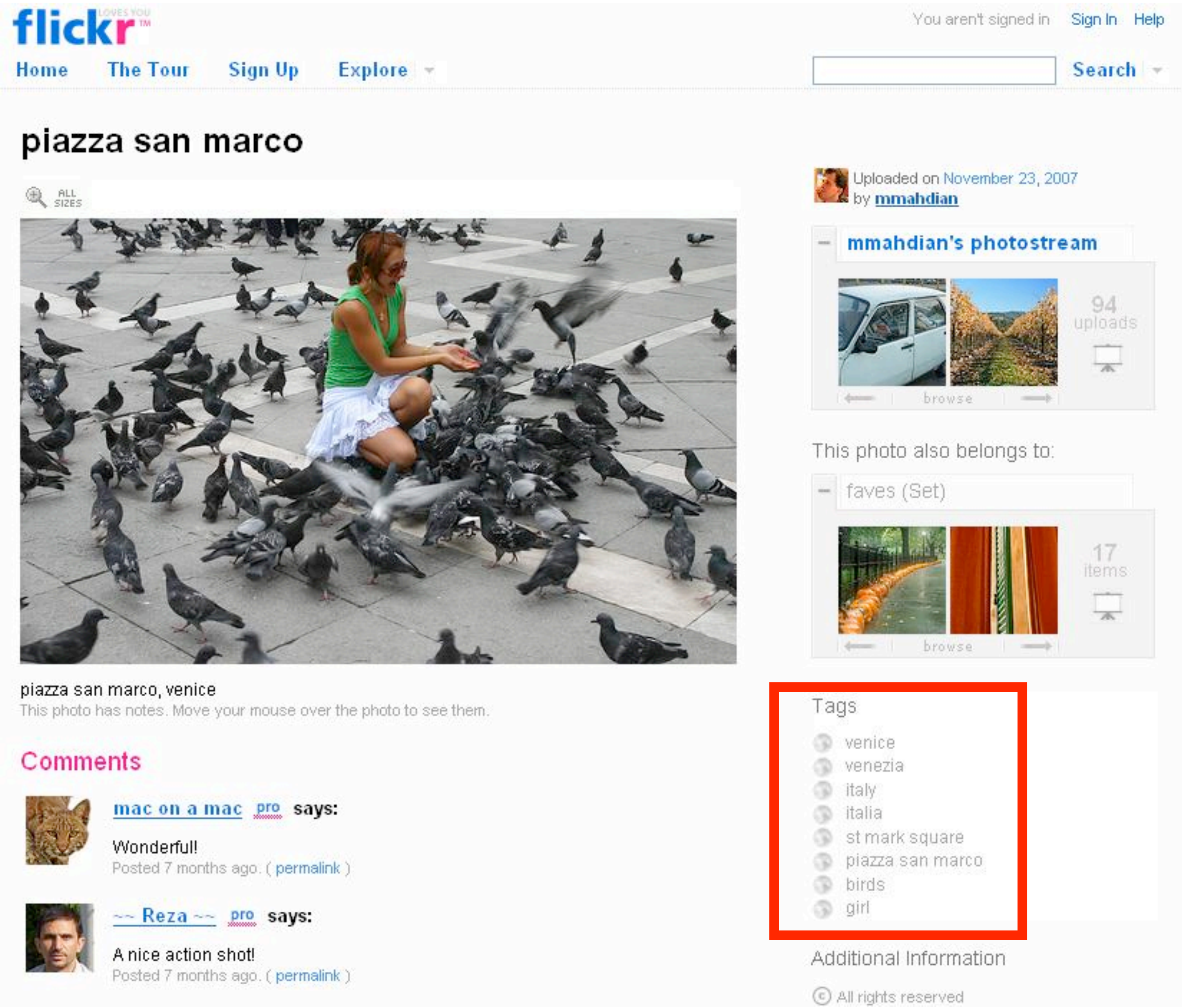
2 comments

**iran**
19 photos

**flowers**
12 photos

**funny pix**
4 photos

**faves**



# piazza san marco

Uploaded on November 23, 2007
by mmahdian

mmahdian's photostream

This photo also belongs to:

faves (Set)

piazza san marco, venice

This photo has notes. Move your mouse over the photo to see them.

Comments

mac on a mac pro says:

Wonderfull!

Posted 7 months ago. ( permalink )

~~ Reza ~~ pro says:

A nice action shot!

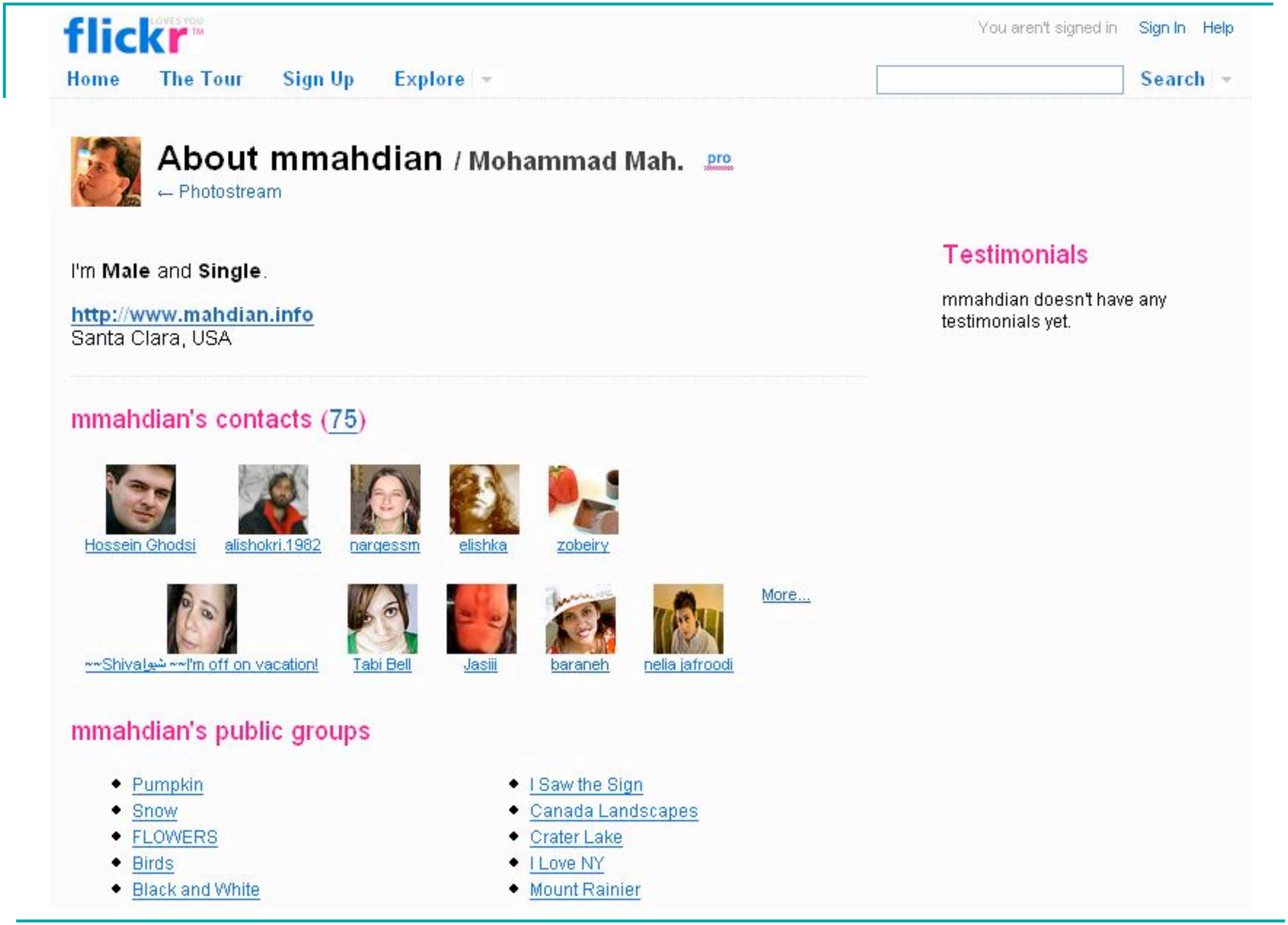
Posted 7 months ago. ( permalink )

Tags

- venice
- venezia
- italy
- italia
- st mark square
- piazza san marco
- birds
- girl

Additional Information

© All rights reserved



# About mmahdian / Mohammad Mah. pro

← Photostream

I'm **Male** and **Single**.

http://www.mahdian.info
Santa Clara, USA

## Testimonials

mmahdian doesn't have any testimonials yet.

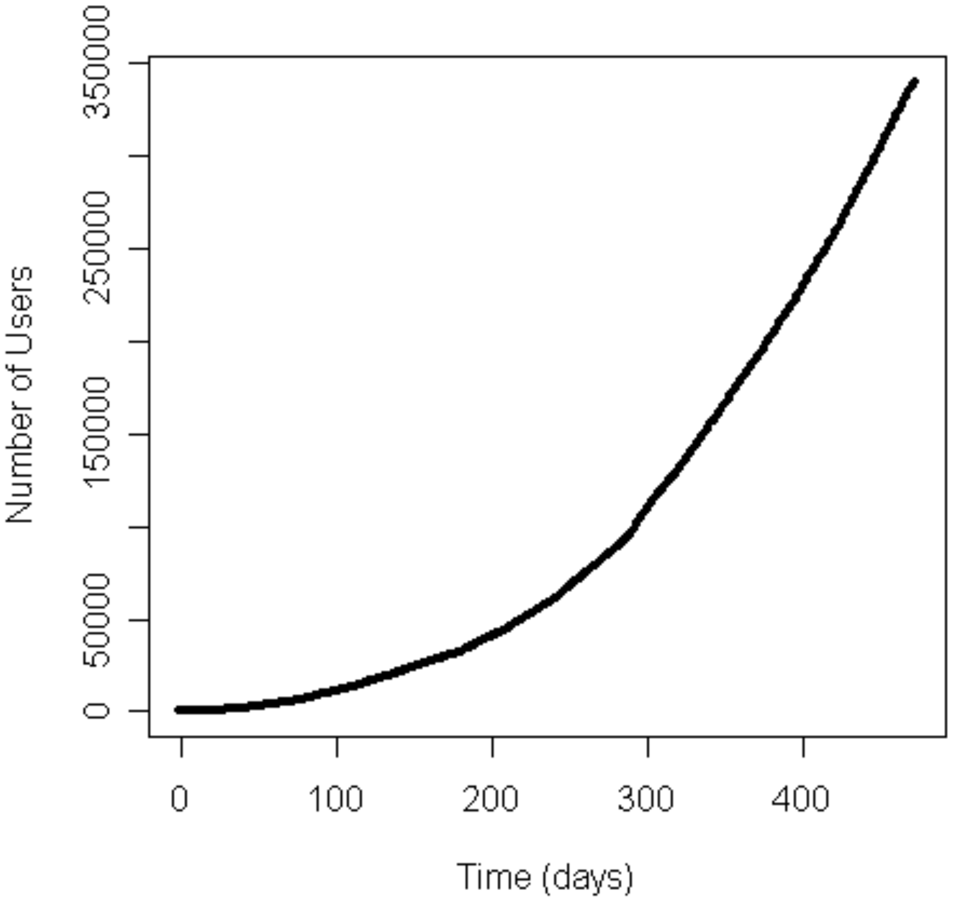
## mmahdian's contacts (75)

Hossein Ghodsi
alishokri.1982
nargessm
elishka
zobeiry
~~Shivalشیوا~~I'm off on vacation!
Tabi Bell
Jasiii
baraneh
nelia jafroodi
More...

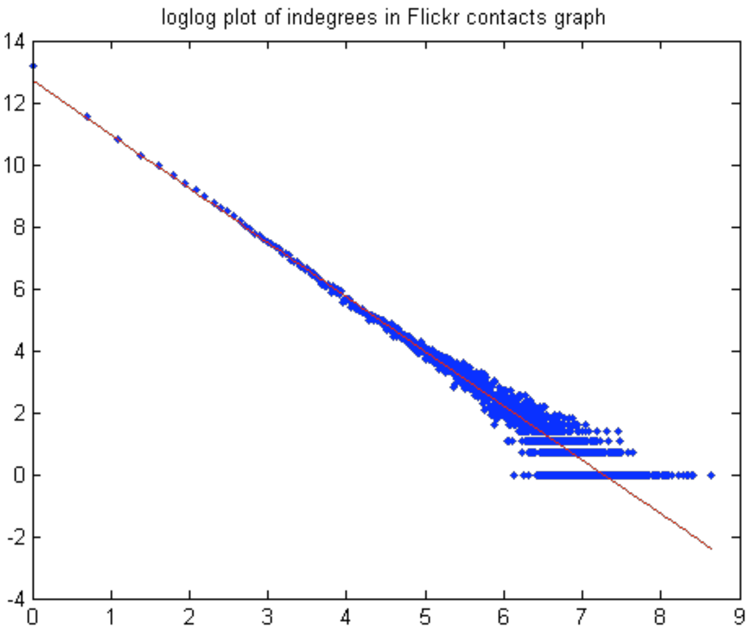
## mmahdian's public groups

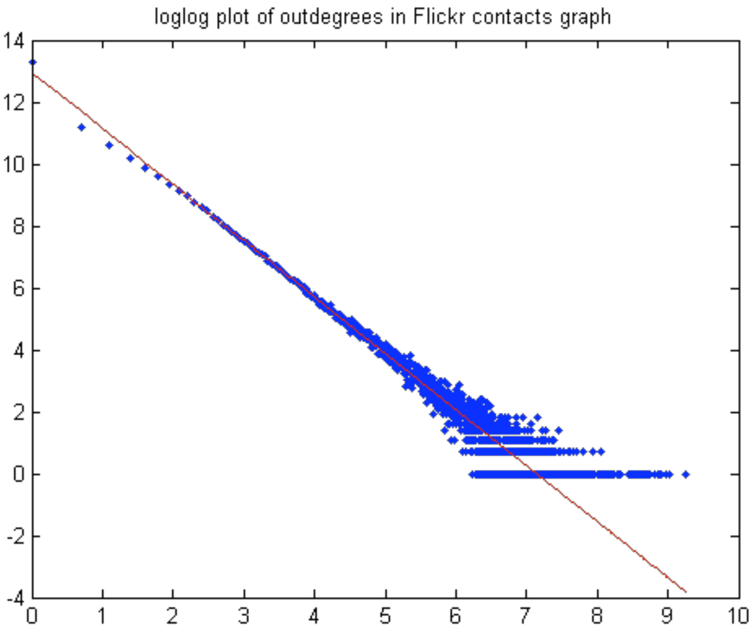
- Pumpkin
- Snow
- FLOWERS
- Birds
- Black and White
- I Saw the Sign
- Canada Landscapes
- Crater Lake
- I Love NY
- Mount Rainier

# Flickr data set, growth



# Flickr graph, indegrees & outdegrees



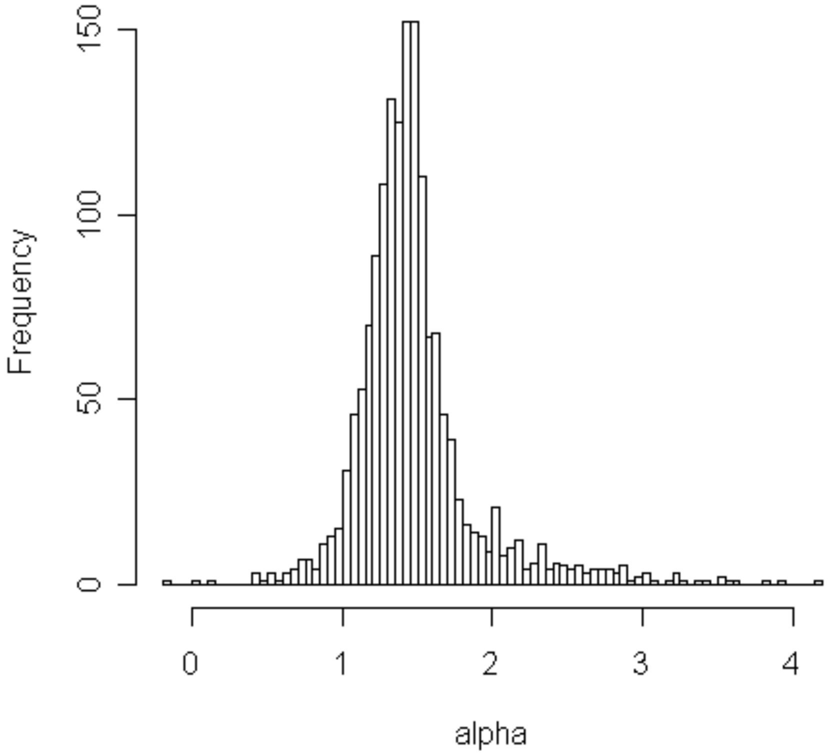


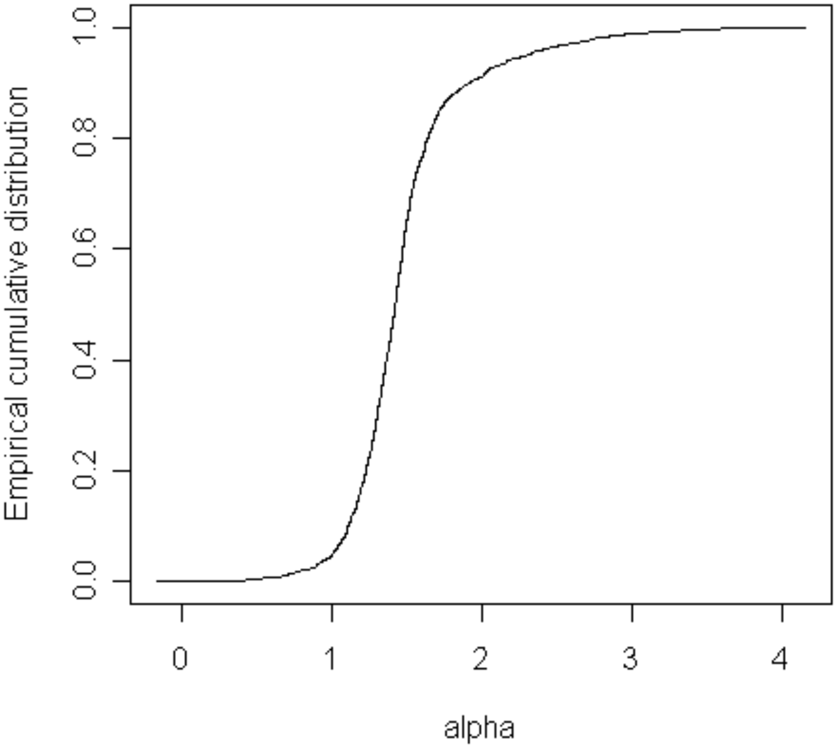
# Flickr tags

- ~10K tags
- We focus on a set of 1700
- Different growth patterns:
    - bursty (“halloween” or “katrina”)
    - smooth (“landscape” or “bw”)
    - periodic (“moon”)
- For each tag, define an action corresponding to using the tag for the first time.

# Social correlation in flickr

- Distribution of ® values estimated using maximum likelihood:





# Distinguishing influence

- Recall: graph G, set W of active nodes
- Non-influence models
  - Homophily: first W is picked, then G is picked from a distribution that depends on W
  - Confounding factors: both G and W are picked from distributions that depend on another var X.
- Generally, we consider this correlation model:
  - (G,W) are selected from a joint distribution
  - Each agent in W picks an activation time i.i.d. from a distribution on [0,T].

# Testing for influence

- Simple idea:  even though an agent’s probability of activation can depend on friends, her timing of activation is independent
- Shuffle Test:   re-shuffle the time-stamp of all actions, and re-estimate the coefficient ®.  If different from original ®, social influence can’t be ruled out.
- Edge-Reversal Test:  reverse the direction of all edges, and re-estimate ®.

# Shuffle Test, Theoretical Justification

- Theorem. If the graph is large enough, time-shuffle test rules out the general model of correlation.

- Intuition: in correlation model, the distribution of the data remains the same if time-stamps are shuffled.

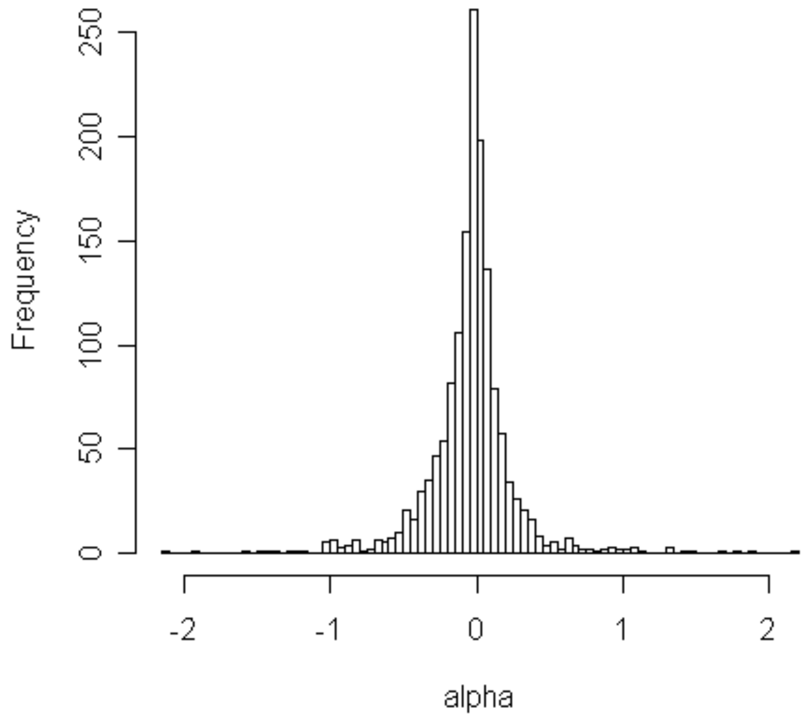
- Challenge: prove concentration.

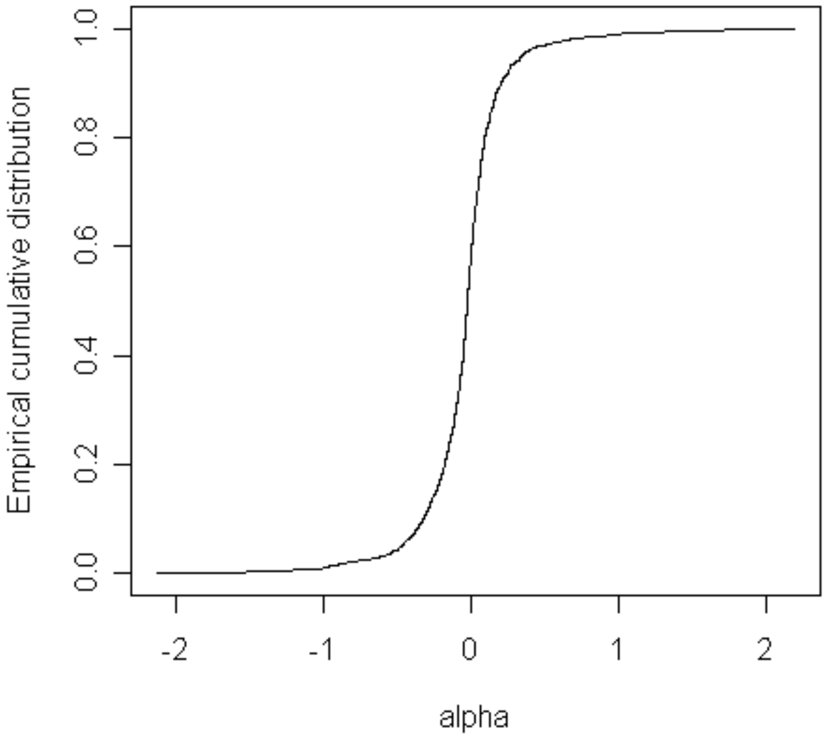
- Proof sketch:
  - First use Azuma's martingale inequality to show that $Y_a$'s and $N_a$'s are concentrated.
  - Then show that the maximum likelihood estimate for ® is a continuous function of $Y_a$'s and $N_a$'s.

# Simulations

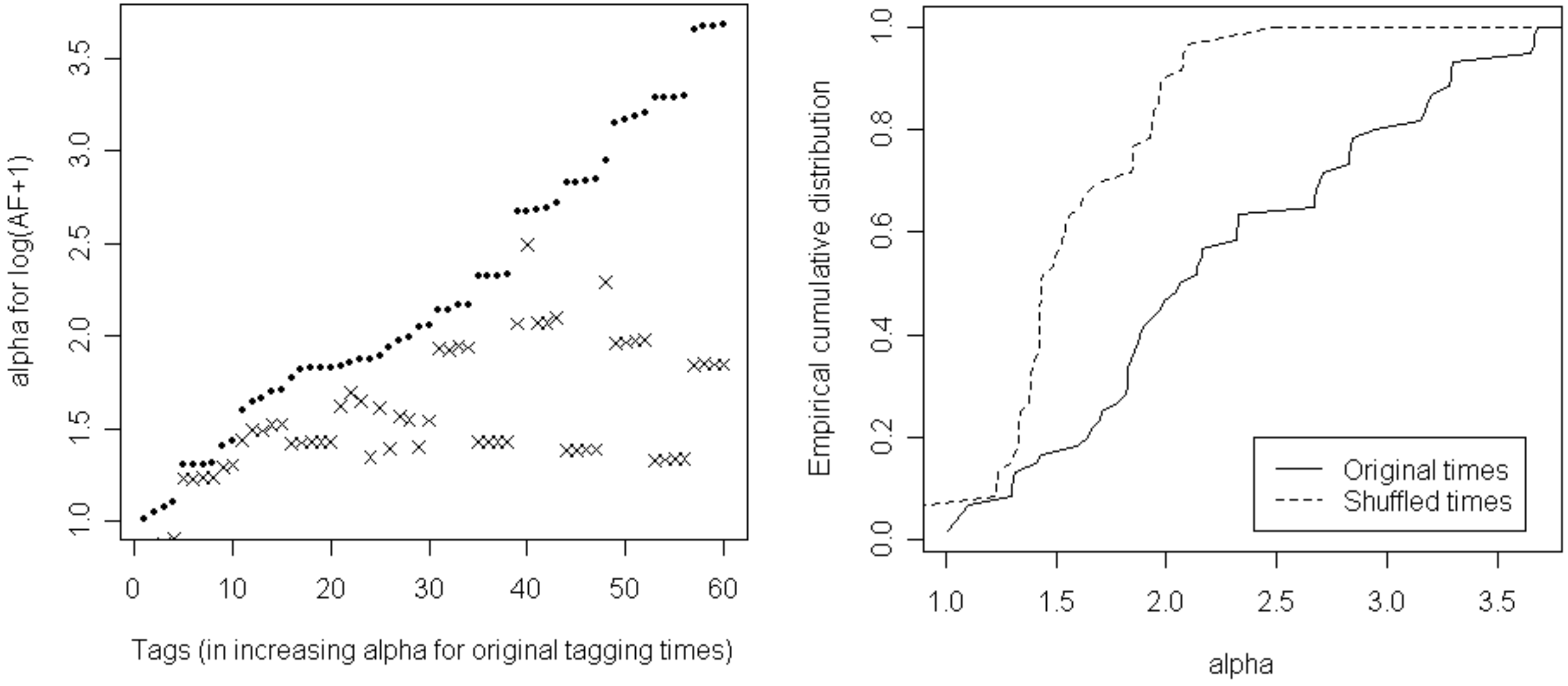
- Run the tests on randomly generated action data on flickr network.
- Baseline: no-correlation model, actions generated randomly to follow the pattern of one of the real tags, but ignoring network
- Influence model: same as described, with a variety of (®,¯) values
- Correlation model: pick a # of random centers, let W be the union of balls of radius 2 around these centers.

# Simulation results, baseline

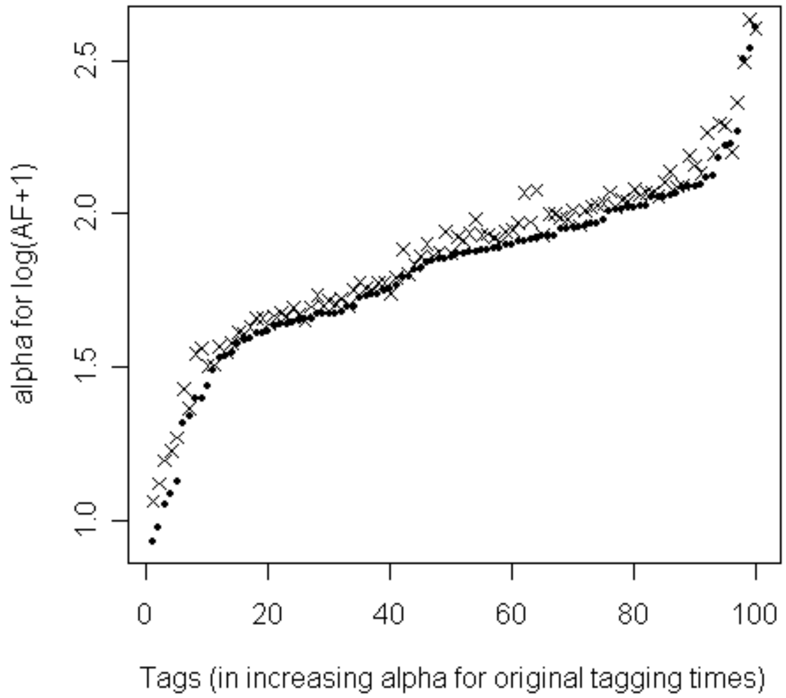


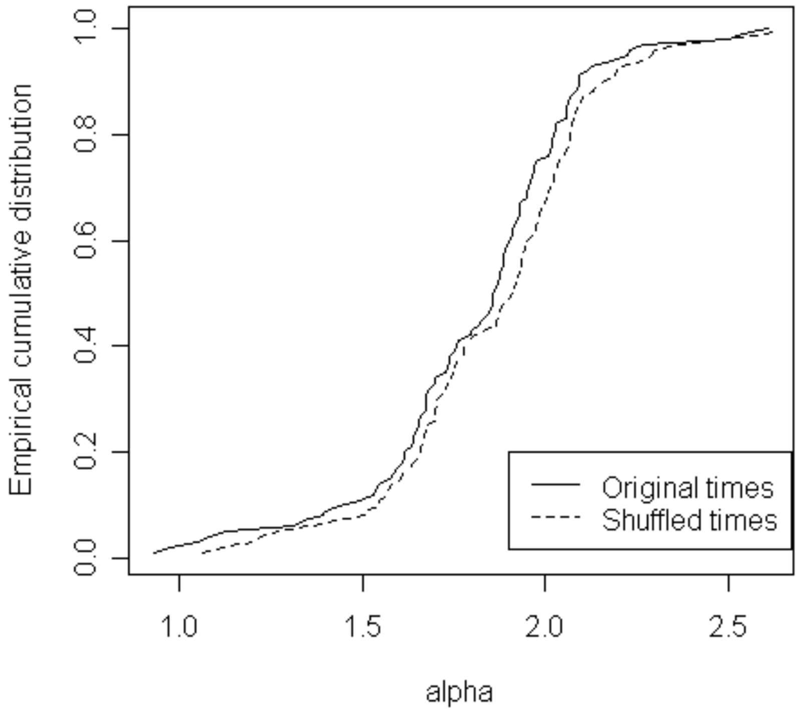


# Shuffle test, influence model

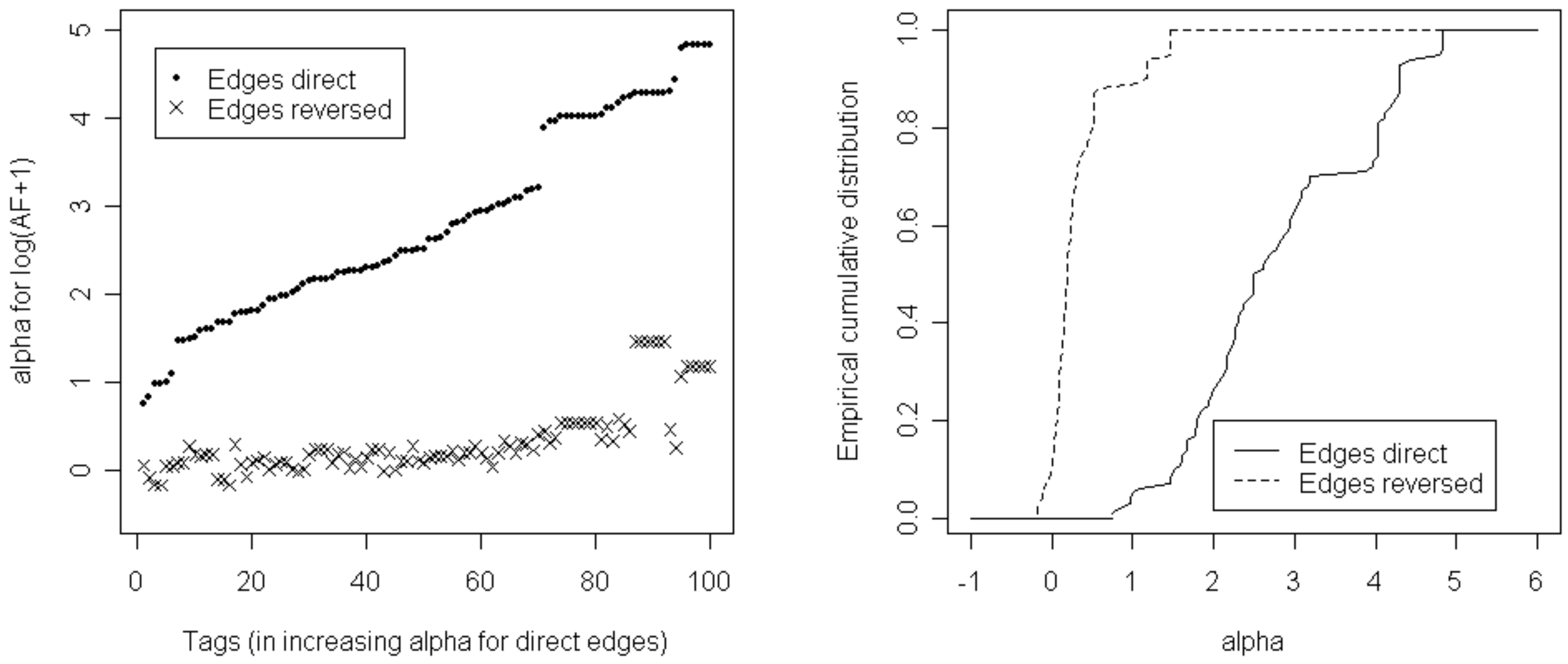


# Shuffle test, correlation model

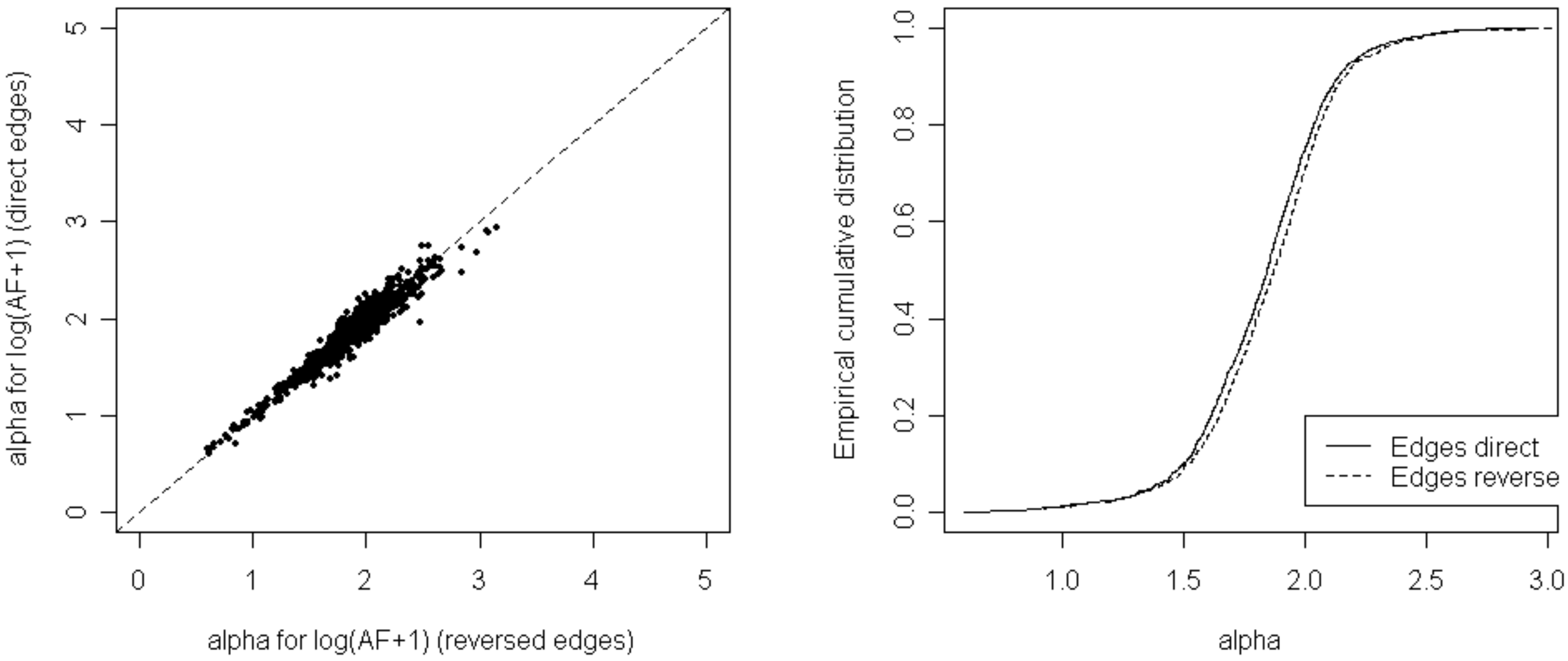




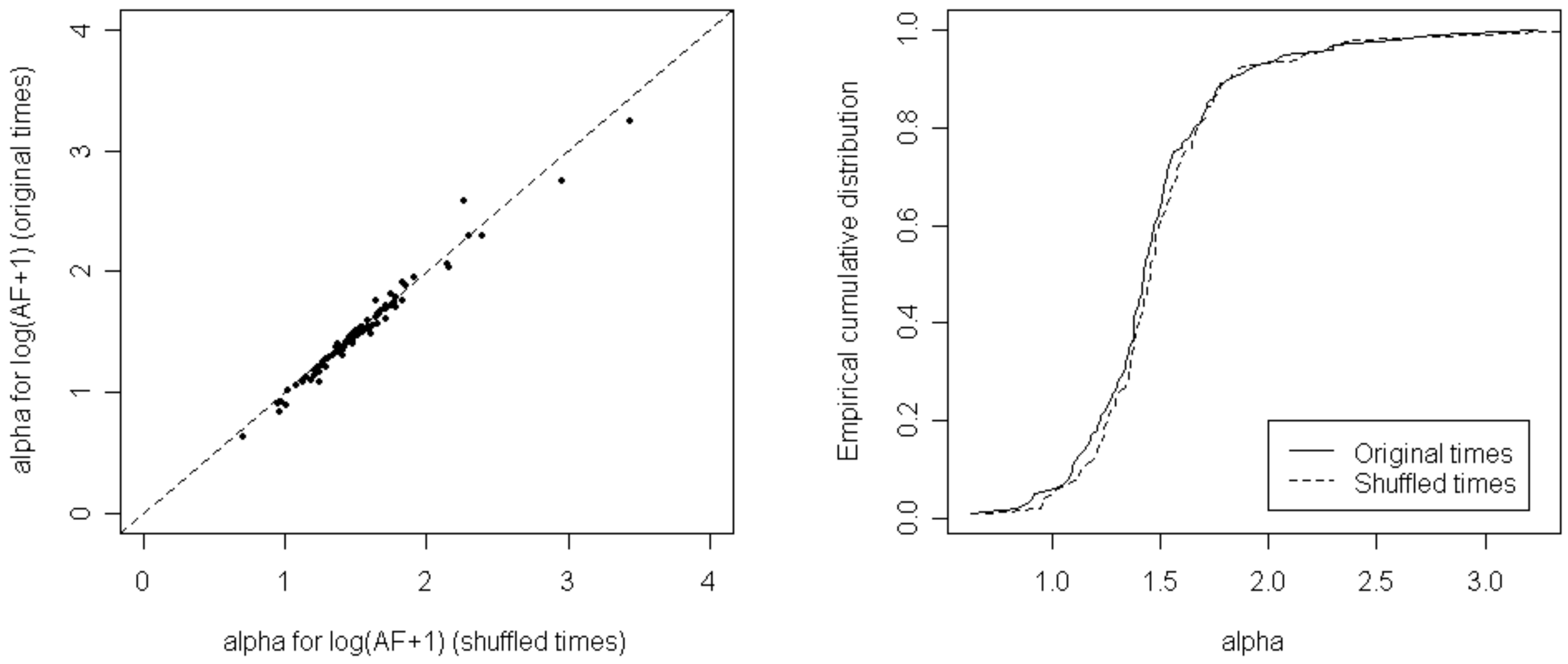
# Edge-reversal test, influence model



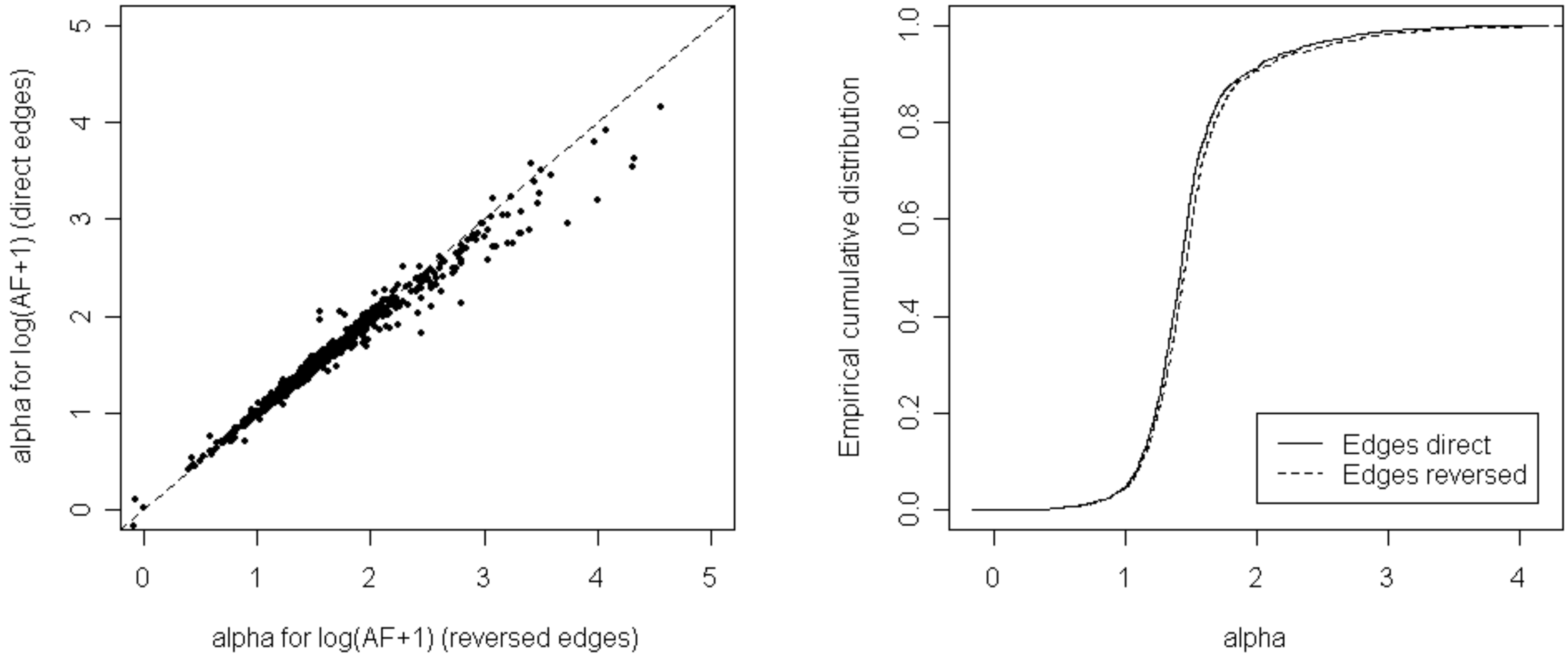
# Edge-reversal test, correlation model



# Shuffle test on Flickr data



# Edge-reversal test on Flickr data



# Results of experiments

- On Flickr, we conclude that despite considerable correlation, no social influence can be detected.
- Discussion
  - cannot conclusively say there is influence without controlled experiments (example: flu shot)
  - still can rule out potential candidates
  - Open: develop algorithms to find "influential" nodes/ communities given a pattern of spread.

# Conclusion

- Social networks are
  - Important subjects of study
  - Useful in understanding dynamics of societies (epidemics, cultural norms, technology adoption, …)
  - Useful for doing things (finding red balloons, citizen journalism, …)
  - To use them, we must have a good understanding of how micro-scale preferences lead to macro-scale phenomena
  - requires algorithmic viewpoint of CS, equilibrium analysis techniques of econ/sociology, modeling techniques of physics, …